Abstract

A technique for extracting a meaningful text block from a document where a table, an itemized list, a multiple column, etc., are arbitrarily laid out. A document is input which is laid out using blanks or the like, then a symbol is acquired which is associated with a spatial coordinate of the document. Consecutive characters of the same type are extracted from the symbol to generate a token and a space. A stream is generated from consecutive spaces in the column direction, while a text block is generated from streams and tokens. A link is generated between the text blocks to form a document graph. Validity of a connection (link) between the text blocks in the document graph is evaluated using a language model, then the text blocks are merged if the connection is valid.